Comput Stat (2015) 30:987–1010
DOI 10.1007/s00180-015-0555-0

ORIGINAL PAPER

A comparison of tests for the one-way ANOVA problem for functional data

Tomasz Górecki · Łukasz Smaga

Received: 17 July 2014 / Accepted: 13 January 2015 / Published online: 29 January 2015

Abstract In this paper, some new tests based on the idea of the B-spline test (Shen and Faraway in Stat Sin 14:1239–1257, 2004) for the one-way ANOVA problem for functional data are proposed. Eleven existing tests for this problem are also reviewed. Exhaustive simulation studies are presented to compare all of the tests considered. The simulations are based on real labeled times series data and artificial data. They provide an idea of the size control and power of the tests, and emphasize the differences between them. Illustrative examples of the use of the tests in practice are also given.

Keywords Basis function representation · Equality of functional means · Longitudinal data · One-way ANOVA for functional data · Orthonormal basis · Stochastic process

1 Introduction

Great advances in computational and analytical techniques now enable many processes to be continuously monitored. To analyze the large quantities of data from such processes, new statistical methods are needed. In functional data analysis (FDA), such data are considered as random functions so-called functional data.

Electronic supplementary material The online version of this article (doi:10.1007/s00180-015-0555-0) contains supplementary material, which is available to authorized users.

T. Górecki · Ł. Smaga (✉)
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Umultowska 87, 61-614 Poznan, Poland
e-mail: ls@amu.edu.pl

T. Górecki
e-mail: tomasz.gorecki@amu.edu.pl

A broad perspective of FDA methods is given in Ramsay and Silverman (2002); Ramsay and Silverman (2005), Ferraty and Vieu (2006), Ramsay et al. (2009), Horváth and Kokoszka (2012), Zhang (2013), and in the review papers Valderrama (2007) and Cuevas (2014). Many standard statistical methods are being adapted for functional data. These include, for example, canonical correlation analysis (Leurgans et al. 1993; Krzyśko and Waszak 2013), cluster analysis (Tokushige et al. 2007; Yamamoto and Terada 2014), discriminant analysis (James and Hastie 2001; Preda et al. 2007; Górecki et al. 2014), principal component analysis (Boente and Fraiman 2000; Berrendero et al. 2011), and regression (Cuevas et al. 2002; Cai and Hall 2006; Benhenni et al. 2007; Chiou and Müller 2007). FDA is also strongly linked with longitudinal data analysis (LDA), which is a better-known area. Both methodologies serve to analyze data collected over time about the same subjects, but they are nonetheless intrinsically different (see Davidian et al. 2004, or Martínez-Camblor and Corral 2011, for more details). The differences between FDA and LDA involve chiefly the viewpoints and ways of thinking about the data. The connection between FDA and LDA is considered in Zhao et al. (2004).

The present paper is concerned with the one-way ANOVA problem for functional data. This problem can be formulated in the following way. Let $X_{i1}(t), X_{i2}(t), \ldots, X_{in_i}(t)$, $i = 1, \ldots, k$ denote k groups of random functions defined over a given finite interval $T = [a, b]$. Let $\mathrm{SP}(m, \gamma)$ denote a stochastic process with mean function $m(t), t \in T$ and covariance function $\gamma(s, t), s, t \in T$. Assuming that $X_{i1}(t), X_{i2}(t), \ldots, X_{in_i}(t)$ are i.i.d. $\mathrm{SP}(m_i, \gamma)$, $i = 1, \ldots, k$, it is often interesting to test the equality of the k mean functions

$$H_0 : m_1(t) = \cdots = m_k(t), \quad t \in T, \tag{1}$$

against the alternative that its negation holds. This problem is known as the k-sample testing problem or the one-way ANOVA problem for functional data. As we will see, some of the tests considered in this paper can also be used in the general "heteroscedastic" case, where the covariance functions in groups are not necessarily equal.

Quite a few tests for the aforementioned problem are given in the literature. Some of these tests have been used in many practical experiments in chemometrics (Bobelyn et al. 2010; Tarrío-Saavedra et al. 2011), economics (Long et al. 2012), transport emissions (Gao 2007), etc. Most of these tests are described briefly in Sect. 2, where we also propose new tests. The tests differ in terms of size control and power. To compare the tests, we present exhaustive simulation studies whose results may help in choosing the best test for a specific real problem.

The rest of the paper is organized as follows. In Sect. 2, the existing tests for the one-way ANOVA problem for functional data are reviewed and new tests for that problem are presented. Simulation studies and illustrative examples are presented in Sects. 3 and 4, respectively. Some conclusions are given in Sect. 5. Supplementary Materials are described in Sect. 6. Proofs of the theoretical results are outlined in the "Appendix".

2 Tests for the one-way ANOVA problem for functional data

In this section, we describe existing tests for the one-way ANOVA problem for functional data and present new tests based on the idea of the B-spline test (Shen and Faraway 2004).

2.1 Existing tests

We first set up notation. Let

$$\mathrm{SSR}_n(t) = \sum_{i=1}^{k} n_i(\bar{X}_i(t) - \bar{X}(t))^2$$

and

$$\mathrm{SSE}_n(t) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}(t) - \bar{X}_i(t))^2$$

denote the pointwise between-subject and within-subject variations respectively, where

$$\bar{X}(t) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}(t)$$

and

$$\bar{X}_i(t) = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}(t),$$

$i = 1, \ldots, k$, are respectively the sample grand mean function and the sample group mean functions. Moreover, let $\mathrm{tr}(\gamma) = \int_T \gamma(t,t)dt$ denote the trace of $\gamma(s,t)$. The pooled sample covariance function $\hat{\gamma}(s,t)$, as an unbiased estimator of $\gamma(s,t)$, is given by

$$\hat{\gamma}(s,t) = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}(s) - \bar{X}_i(s))(X_{ij}(t) - \bar{X}_i(t)). \tag{2}$$

The natural extension of the classical ANOVA F-test for real variables to the context of functional data analysis is the pointwise F-test proposed by Ramsay and Silverman (2005, p. 227). The test statistic of the pointwise F-test for (1) is defined as

$$F_n(t) = \frac{\mathrm{SSR}_n(t)}{k-1} \Big/ \frac{\mathrm{SSE}_n(t)}{n-k}.$$

The pointwise F-test rejects (1) at each $t \in T$ whenever $F_n(t) > F_{k-1,n-k}(1-\alpha)$ for any predetermined significant level α, where $F_{k-1,n-k}(1-\alpha)$ denotes the upper $100(1-\alpha)$ percentile of $F_{k-1,n-k}$. Hence we can test (1) at all points of T using the same critical value for any predetermined significant level. Unfortunately, the pointwise F-test is time-consuming (it must be performed for all $t \in T$) and it is not guaranteed that the one-way ANOVA problem (1) is significant overall for a given significance level even when the pointwise F-test is significant for all $t \in T$ at the same significance level. In the literature, global tests which overcome this difficulty are also proposed. We briefly describe them below.

Cuevas et al. (2004) proposed to use the following test statistic for testing (1)

$$V_n = \sum_{1 \le i < j \le k} n_i \int_T (\bar{X}_i(t) - \bar{X}_j(t))^2 dt.$$

Under the null hypothesis (1) and the assumptions that $n_i, n \to \infty$ in such a way that $n_i/n \to p_i > 0$ for $i = 1, \ldots, k$, they proved that the approximate distribution of V_n is that of the statistic

$$V = \sum_{1 \le i < j \le k} n_i \int_T (Z_i(t) - \sqrt{p_i/p_j} Z_j(t))^2 dt, \tag{3}$$

where $Z_1(t), \ldots, Z_k(t)$ are independent Gaussian processes with mean 0 and covariance function $\gamma(s,t)$. Cuevas et al. (2004) computed the p-value of V_n, or its empirical critical value, by resampling $Z_i(t)$, $i = 1, \ldots, k$, from Gaussian processes GP$(0, \hat{\gamma})$, where $\hat{\gamma}(s,t)$ is given by (2), a large number of times. This test will be referred to as the CH test. They also showed that the test statistic V_n can be used for testing (1) in the "heteroscedastic" case. In this case, the p-value of V_n can be computed as above, but here $Z_1(t), \ldots, Z_k(t)$ are independent Gaussian processes with covariance functions

$$\hat{\gamma}_i(s,t) = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij}(s) - \bar{X}_i(s))(X_{ij}(t) - \bar{X}_i(t)), \quad i = 1, \ldots, k.$$

This version of the test of Cuevas et al. (2004) will be referred to as the CS test.

The L^2-norm-based test (Faraway 1997; Zhang and Chen 2007) adopted for (1) uses the test statistic $S_n = \int_T \mathrm{SSR}_n(t) dt$. Under the null hypothesis (1), it can be shown that $S_n \sim \beta \chi^2_d$ approximately, where $\beta = \mathrm{tr}(\gamma^{\otimes 2})/\mathrm{tr}(\gamma)$, $d = (k-1)\kappa$, $\kappa = \mathrm{tr}^2(\gamma)/\mathrm{tr}(\gamma^{\otimes 2})$ and $\gamma^{\otimes 2}(s,t) = \int_T \gamma(s,u)\gamma(u,t) du$. This approximate distribution can be used to compute the p-value of S_n ($P(\chi^2_d \ge S_n/\beta)$) or its critical value ($\beta \chi^2_d (1-\alpha)$). In practice, the parameters β and κ are estimated based on the functional data by a naive method or a bias-reduced method. With the estimator $\hat{\gamma}(s,t)$ given in (2), by the naive method, we have $\hat{\beta} = \mathrm{tr}(\hat{\gamma}^{\otimes 2})/\mathrm{tr}(\hat{\gamma})$, $\hat{d} = (k-1)\hat{\kappa}$ and $\hat{\kappa} = \mathrm{tr}^2(\hat{\gamma})/\mathrm{tr}(\hat{\gamma}^{\otimes 2})$, and by the bias-reduced method, we have $\hat{\beta} = \widehat{\mathrm{tr}(\gamma^{\otimes 2})}/\mathrm{tr}(\hat{\gamma})$, $\hat{d} = (k-1)\hat{\kappa}$ and $\hat{\kappa} = \widehat{\mathrm{tr}^2(\gamma)}/\widehat{\mathrm{tr}(\gamma^{\otimes 2})}$, where

$$\widehat{\mathrm{tr}(\gamma^{\otimes 2})} = \frac{(n-k)^2}{(n-k-1)(n-k+2)}\left(\mathrm{tr}(\hat{\gamma}^{\otimes 2}) - \frac{\mathrm{tr}^2(\hat{\gamma})}{n-k}\right),$$
$$\widehat{\mathrm{tr}^2(\gamma)} = \frac{(n-k)(n-k+1)}{(n-k-1)(n-k+2)}\left(\mathrm{tr}^2(\hat{\gamma}) - \frac{2\mathrm{tr}(\hat{\gamma}^{\otimes 2})}{n-k+1}\right).$$

Therefore, we have two L^2-norm-based tests which differ in the method of estimation of the parameters β and κ, i.e. the L^2-norm-based test with the naive method of estimation of those parameters (the $\mathrm{L}^2\mathrm{N}$ test for short), and the L^2-norm-based test with the bias-reduced method of estimation of those parameters (the $\mathrm{L}^2\mathrm{B}$ test for short).

The F-type test (Shen and Faraway 2004; Zhang 2011) adopted for (1) uses the test statistic

$$F_n = \frac{\int_T \mathrm{SSR}_n(t)dt/(k-1)}{\int_T \mathrm{SSE}_n(t)dt/(n-k)}.$$

Under the null hypothesis (1), we can show that $F_n \sim F_{d_1,d_2}$ approximately, where d_1 is the same as for the L^2-norm-based test given earlier and $d_2 = (n-k)\kappa$. Similarly, the approximate null distribution of the F-type test can be used to compute the p-value of F_n or its critical value. The parameter κ can be estimated by the naive method or by the bias-reduced method described above. Hence we consider the F-type test with the naive estimation method for the parameter κ (the FN test for short), and the F-type test with the bias-reduced estimation method for the parameter κ (the FB test for short). Further details about the aforementioned L^2-norm-based and F-type tests for one-way ANOVA for functional data can be found in Zhang (2013, ch. 5).

When the k samples are not Gaussian and when the sample sizes are small, the L^2-norm-based test and the F-type test are not preferred (see Zhang 2013). In this case, the bootstrap versions of these tests can be used to bootstrap the p-values of S_n and F_n (see Zhang 2013, p. 159, for more details). To shorten the notation, we refer to the L^2-norm-based bootstrap test and F-type bootstrap test as the $\mathrm{L}^2\mathrm{b}$ test and Fb test, respectively.

Globalization of the pointwise F-test (the GPF test, Zhang and Liang 2014) uses the test statistic $T_n = \int_T F_n(t)dt$. Under the null hypothesis (1), it can be shown that $T_n \sim \hat{\beta}_w \chi^2_{\hat{d}_w}$ approximately, where

$$\hat{\beta}_w = (n-k-2)\mathrm{tr}(\hat{\gamma}_w^{\otimes 2})/((k-1)(n-k)(b-a)),$$
$$\hat{d}_w = (k-1)(n-k)^2(b-a)^2/((n-k-2)^2\mathrm{tr}(\hat{\gamma}_w^{\otimes 2}))$$

and

$$\hat{\gamma}_w(s,t) = \hat{\gamma}(s,t)/\sqrt{\hat{\gamma}(s,s)\hat{\gamma}(t,t)},$$

where $\hat{\gamma}(s,t)$ is given by (2). This approximate distribution can be used to compute the p-value of T_n or its critical value in much the same way as for the L^2-norm-based test.

Fan and Lin (1998) proposed to use adaptive Neyman test to compare different sets of curves. This test can be adopted for testing (1) (see, for example, Laukaitis and Račkauskas 2005, which used it in the analysis for clients segmentation tasks). Let $X^*_{ij}(l), l = 1, \ldots, T^*$ be the discrete Fourier tranform of the function X_{ij} for given i and j. We assume the model $X^*_{ij}(l) = m^*_i(l) + \epsilon^*_{ij}(l)$, where $\epsilon^*_{ij}(l)$ are independent and $\epsilon^*_{ij}(l) \sim N(0, \sigma^2_i(l))$. The null hypothesis (1) is transformed then to $H^*_0 : m^*_1(l) = \cdots = m^*_k(l) = m^*(l)$ for all l. To test H^*_0, Fan and Lin (1998) propose to use the test statistic

$$T_{\text{HANOVA}} = \sqrt{2 \log \log T^*} F_{\hat{m}} - \{2 \log \log T^* + 0.5 \log \log \log T^* - 0.5 \log(4\pi)\},$$

where

$$F_{\hat{m}} = \max_{1 \leq m \leq T^*} \frac{1}{\sqrt{2(k-1)m}} \left[\sum_{l=1}^{m} \sum_{i=1}^{k} n_i \hat{\sigma}^*_i(l)^{-2} [\bar{X}^*_i(l) - \bar{X}^*(l)]^2 - (k-1)m \right],$$

and

$$\bar{X}^*_i(l) = \frac{1}{n_i} \sum_{j=1}^{n_i} X^*_{ij}(l), \quad \hat{\sigma}^*_i(l) = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X^*_{ij}(l) - \bar{X}^*_i(l))^2,$$

$$\bar{X}^*(l) = \sum_{i=1}^{k} n_i \hat{\sigma}^*_i(l)^{-2} \bar{X}^*_i(l) \Big/ \sum_{i=1}^{k} n_i \hat{\sigma}^*_i(l)^{-2}.$$

As it is discussed in Fan and Lin (1998), the choice of the maximum number of dimensions T^* does not alter the result very much, as long as T^* is large enough so that high-frequency cells are basically noise. We reject H^*_0 when T_{HANOVA} is too large. The critical values for this test statistic can be obtained by the finite sample distribution of it (see Table 1 in Fan and Lin 1998). We refer to this test as the FL test.

2.2 Tests based on a basis function representation

In this section, we present tests for the one-way ANOVA problem which are based on a basis function representation of the stochastic processes $(X_{ij}(t), t \in T)$. These tests are inspired by the idea of the B-spline method of Shen and Faraway (2004).

Assume that we observe k groups of stochastic processes $X_{ij} \in L_2(T)$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, where $L_2(T)$ is the Hilbert space of square integrable functions on the interval T, equipped with the inner product $< f, g > = \int_T f(t)g(t)dt$. Let $\{\phi_l\}$ be an orthonormal basis of $L_2(T)$. This orthonormal basis also called basis function system has the property that we can approximate arbitrarily well any function by taking a weighted sum or linear combination of a sufficiently large number K of these functions (see Ramsay and Silverman 2005). For this reason, we consider the case where the stochastic processes $(X_{ij}(t), t \in T)$ can be represented by a finite number of orthonormal basis functions, i.e.

$$X_{ij}(t) = \sum_{l=0}^{K} c_{ijl}\phi_l(t), \quad t \in T, \tag{4}$$

where c_{ijl}, $l = 0, 1, \ldots, K$, are random variables with finite variance and K is sufficiently large. We present how we can choose the appropriate value of K based on the data later on.

We write $\boldsymbol{\phi}(t) = (\phi_0(t), \phi_1(t), \ldots, \phi_K(t))'$ and $\mathbf{c}_{ij} = (c_{ij0}, c_{ij1}, \ldots, c_{ijK})'$. The stochastic processes $(X_{ij}(t), t \in T)$, the sample grand mean function and the sample group mean functions can be written in matrix notation as follows:

$$X_{ij}(t) = \mathbf{c}_{ij}'\boldsymbol{\phi}(t), \quad \bar{X}(t) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{c}_{ij}'\boldsymbol{\phi}(t), \quad \bar{X}_i(t) = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{c}_{ij}'\boldsymbol{\phi}(t), \tag{5}$$

for $t \in T$.

The vectors $\mathbf{c}_{ij}$ can be estimated by the least squares method, and the optimum value of K in the sense of smoothness can be selected for each process $(X_{ij}(t), t \in T)$ using the Bayesian Information Criterion (BIC); then from the values of K corresponding to all processes a modal value is selected as the common value for all $(X_{ij}(t), t \in T)$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$ (see Krzyśko and Waszak 2013; Górecki et al. 2014, for more details). We should prefer K to be large, particularly when the stochastic processes $(X_{ij}(t), t \in T)$ are observed at high frequency with little noise.

We reduce the stochastic processes $(X_{ij}(t), t \in T)$ to vectors $\mathbf{c}_{ij}$ of length $K + 1$, for all $i = 1, \ldots, k$, $j = 1, \ldots, n_i$. Of course, the score vectors $\mathbf{c}_{ij}$ contain some information about these processes. The B-spline method in Shen and Faraway (2004) for testing hypotheses for functional data uses this information in the following simple way: We represent each stochastic process as a linear combination of orthonormal basis functions and then perform the usual multivariate test on the coefficients of this representation. Based on this idea, the usual MANOVA tests on the vectors $\mathbf{c}_{ij}$ can be performed to test (1). The well-known MANOVA tests are Wilk's lambda (which is a function of the likelihood ratio test statistic), the Lawley-Hotelling trace, the Pillai trace, and Roy's maximum root (Anderson 2003). Therefore, we have four tests for the one-way ANOVA problem for functional data, and we call these the W, LH, P and R tests, from the initial letters of the surnames of their originators.

The aforementioned tests for the one-way ANOVA problem for functional data are very simple, but we will see that they do not perform so well as the other methods in some cases. Moreover, from the formal requirements of the MANOVA tests, we usually have to reduce the number of basis functions in (4), which can negatively affect the smoothness of a basis functional representation of the processes $(X_{ij}(t), t \in T)$. However, a basis function representation can be used to construct a better test for (1). The test statistic of the classical ANOVA F-test for real variables adopted for functional data takes the form

$$F = \frac{\frac{1}{k-1}\sum_{i=1}^{k} n_i ||\bar{X}_i - \bar{X}||_2^2}{\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} ||X_{ij} - \bar{X}_i||_2^2}, \tag{6}$$

where $||f||_2^2 = \int_T f^2(t)dt$ for $f \in L_2(T)$. In the following proposition, a more useful form of this statistic is proved.

Proposition 1 *Assume that the stochastic processes $X_{ij} \in L_2(T)$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, are represented by a finite number of orthonormal basis functions, i.e. the equations in (5) hold with a known (fixed) K. Then the statistic F given by (6) is equal to*

$$\frac{\frac{1}{k-1}(a-b)}{\frac{1}{n-k}(c-a)},$$

where

$$a = \sum_{i=1}^{k} \frac{1}{n_i} \sum_{m=1}^{n_i} \sum_{s=1}^{n_i} \mathbf{c}'_{im}\mathbf{c}_{is}, \quad b = \frac{1}{n} \sum_{i=1}^{k} \sum_{m=1}^{n_i} \sum_{t=1}^{k} \sum_{v=1}^{n_t} \mathbf{c}'_{im}\mathbf{c}_{tv}, \quad c = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbf{c}'_{ij}\mathbf{c}_{ij}.$$

The proof of Proposition 1 is given in the "Appendix". We propose to use F as the test statistic and a permutation-based p-value for testing (1). Proposition 1 implies that the statistic F given by (6) is a function of the coefficient vectors $\mathbf{c}_{ij}$ and does not depend on the functions ϕ_l. Moreover, it is easy to see that any permutation of the stochastic processes $(X_{ij}(t), t \in T)$ leaves the values of the sums b and c unchanged. Hence for each random permutation of the data only the sum a has to be calculated. Furthermore, the sums a, b and c given in Proposition 1 can be expressed in the forms presented in the following lemma, which are easier and faster to compute, using computer programs such as R, than the sums given in Proposition 1.

Lemma 1 *Let $\mathbf{C}_i = (\mathbf{c}_{i1}, \ldots, \mathbf{c}_{in_i})$, $i = 1, \ldots, k$, where the vectors $\mathbf{c}_{ij}$, $j = 1, \ldots, n_i$, are defined in (5). Then*

$$a = \sum_{i=1}^{k} \frac{1}{n_i} \mathbf{1}'_{n_i} \mathbf{C}'_i \mathbf{C}_i \mathbf{1}_{n_i}, \quad b = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{k} \mathbf{1}'_{n_i} \mathbf{C}'_i \mathbf{C}_j \mathbf{1}_{n_j}, \quad c = \sum_{i=1}^{k} \operatorname{tr}(\mathbf{C}'_i \mathbf{C}_i),$$

where $\mathbf{1}_p$ is the $p \times 1$ vector of ones.

The proof of Lemma 1 is outlined in the "Appendix". The permutation test based on the test statistic F and implemented with the use of the above facts is comparable in terms of speed to the bootstrap tests described in Sect. 2.1. Moreover, as we will see later, this test is comparable in terms of size control and power to other tests for (1), and it performs better than those tests when the number of observations and number of time points are small. We also remark that it is easy to see that $F = F_n$, where F_n is the test statistic of the F-type test described in Sect. 2.1. However, from the simulations in Sect. 3, we find that the above permutation test based on the test statistic F (the FP test for short) performs better than the F-type tests based on approximation of the distribution of this test statistic, and also better than the Fb test.

2.3 Numerical implementation of the tests

In practice, the k functional samples are not continuously observed. Each function is usually observed on a grid of design time points. In this paper, all individual functions in the simulations and the examples are observed on a common grid of design time points. All tests can be directly applied to such functional data. In the Supplementary Materials, we present and describe the R codes which perform all of the global tests considered in this paper. However, the design time points may in some situations be different for different individual functions. To implement the tests numerically for such situations, one first has to reconstruct the k functional samples from the k observed discrete functional samples using some smoothing technique, then discretize each individual function of the k reconstructed functional samples on a common grid of time points, and finally apply the tests accordingly (see Zhang 2013, or Zhang and Liang 2014, for more details).

In all simulations and examples given in this paper, the p-values of the FP test were obtained on the basis of 1,000 permutation replicates, the p-values of the CH and CS tests were evaluated on the basis of 2,000 discretized artificial trajectories for each process $Z_i(t)$ appearing in the limit statistic V given in (3), and the p-values of the L^2b and Fb tests were obtained on the basis of 10,000 bootstrap replicates. The orthonormal basis functions chosen in performing the FP, LH, R, P and W tests are the Fourier system (see Krzyśko and Waszak 2013). The optimal values of K as given in (4) were selected using the BIC from the set $\mathcal{K} = \{3, 5, \ldots, 101\}$ in the case of the FP test. For the LH, R, P and W tests, we usually selected the optimal values of K from a subset of that set whose elements enable the performance of the MANOVA tests. The set $\mathcal{K}$ is associated with the function in the R program which creates the Fourier system. The critical values for the FL test statistic (0.05 upper quantiles of the distribution of T_{HANOVA}) were obtained by the finite sample distribution of it based on one million simulations (see Table 1 in the Supplementary Materials). The R code of the program, which generates these critical values is also given there. The maximum number of dimensions T^* in T_{HANOVA} is chosen in the following way: Suppose that the number of design time points on which the functions are observed is T. If $T < 100$, $T \in [100, 200]$, $T > 200$, then $T^* = T$, $T^* = \lceil T/2 \rceil$, $T^* = 100$, respectively.

3 Simulation studies

In this section we present some simulation studies which serve to compare the tests for the one-way ANOVA problem for functional data. The simulation studies are based on real labeled times series data, which are in fact discrete functional data. They are labeled, which means that the assignment of the observations to groups is known. They also consist of real data. Additional simulation studies based on artificial data are also given.

3.1 Experimental setup

We use the labeled time series data to generate $k = 3$ discrete functional samples with $n_1 = n_2 = n_3 = \bar{n} = 10, 20, 30$. Information on the time series used is given in

Table 1 Summary of data sets

Data set	Number of classes	Number of observations	Time series length
Adiac	37	781	176
CBF	3	930	128
ChlorineConcentration	3	4,307	166
CinC_ECG_torso	4	1,420	1,639
Cricket_X	12	780	300
Cricket_Y	12	780	300
DiatomSizeReduction	4	322	345
Face (all)	14	2,250	131
Face (four)	4	112	350
Fish	7	350	463
Haptics	5	463	1,092
Mallat	8	2,400	1,024
MedicalImages	10	1,141	99
Non-Invasive Thorax1	42	3,765	750
Non-Invasive Thorax2	42	3,765	750
OSU Leaf	6	442	427
Plane	7	210	144
StarLightCurves	3	9,236	1,024
Swedish Leaf	15	1,125	128
Symbols	6	1,020	398
Synthetic Control	6	600	60
Trace	4	200	275
uWaveGestureLibrary_X	8	4,478	315
uWaveGestureLibrary_Y	8	4,478	315
uWaveGestureLibrary_Z	8	4,478	315

Table 1. The time series originate from the UCR Time Series Classification/Clustering Homepage (Keogh et al. 2011), which includes the majority of all of the world's publicly available, labeled time series data sets. The data sets originate from a large number of different domains, including medicine, robotics, astronomy, biology, face recognition, handwriting recognition, etc.

Assume that the classes of the data sets are numbered from 1 to l. Of course, the number of classes l depends on the data set (see Table 1). For each data set described in Table 1 except data sets where there is not a sufficient number of observations, the $k = 3$ functional samples are generated to compare the empirical sizes of the tests for (1) in the following way: (1) select randomly one element i from the set $\{1, \ldots, l\}$; (2) from the ith class, select randomly $3\bar{n}$ observations and create from them three samples, each of size $\bar{n}$. For such samples, we expect to accept the null hypothesis (1). For each data set, we generate 100 such samples. For each of them, the p-values provided by all considered tests for (1) were noted. The rates of rejection of the null

hypothesis (1) at the significance level $\alpha = 0.05$ for all tests are given in Table 2 and Tables 2 and 3 in the Supplementary Materials. This simulation will be referred to as S1.

For each data set described in Table 1 except data sets where there is not a sufficient number of observations, the $k = 3$ functional samples are generated to compare the empirical powers of the tests for (1) in the following way: (1) select randomly two different elements i and j from the set $\{1, \ldots, l\}$; (2) from the ith class, select randomly $2\bar{n}$ observations and create from them two samples, each of size $\bar{n}$; (3) from the jth class, select randomly $\bar{n}$ observations to create the third sample. For the samples generated in this way we expect to reject the null hypothesis (1). The rates of rejection of the null hypothesis (1) at the significance level $\alpha = 0.05$ are given in Table 3 and Tables 4 and 5 in the Supplementary Materials. This simulation will be called S2.

The aforementioned simulations are based on time series whose lengths are mainly moderate or large. We also give additional simulation studies to present the differences between empirical sizes and powers of the tests for the one-way ANOVA problem for "short" functional data. Similarly to Cuevas et al. (2004), we considered an artificial example with $[a, b] = [0, 1]$ and $k = 3$ groups in seven cases:

(M1) $m_i(t) = t(1-t)$ for $i = 1, 2, 3$,
(M2) $m_i(t) = 0.1|\sin(4\pi t)|$ for $i = 1, 2, 3$,
(M3) $m_i(t) = t^i(1-t)^{6-i}$ for $i = 1, 2, 3$,
(M4) $m_i(t) = t^{i/5}(1-t)^{6-i/5}$ for $i = 1, 2, 3$,
(M5) $m_1(t) = 0.05|\sin(4\pi t)|$, $m_2(t) = 0.1|\sin(4\pi t)|$, $m_3(t) = 0.15|\sin(4\pi t)|$,
(M6) $m_1(t) = 0.025|\sin(4\pi t)|$, $m_2(t) = 0.05|\sin(4\pi t)|$, $m_3(t) = 0.075|\sin(4\pi t)|$,
(M7) $m_i(t) = 1 + i/50$ for $i = 1, 2, 3$.

Cases M1 and M2 correspond to situations where H_0 is true; M3–M7 correspond to situations where H_0 is false. For each choice M1–M7 of mean functions, we generated independent samples with $n_1 = n_2 = n_3 = \bar{n} = 10, 20, 30$, under the model $X_{ij}(t) = m_i(t) + \epsilon_{ij}(t)$, $i = 1, 2, 3$, $j = 1, \ldots, \bar{n}$. More precisely, the processes $(X_{ij}(t), t \in [0, 1])$ were generated in discretized versions $X_{ij}(t_s)$, for $s = 1, \ldots, 25$, where the values t_s were chosen equispaced in the interval $[0, 1]$. Two different type of errors are considered. In the *normal* case for the models M1–M7, the errors $\epsilon_{ij}(t_s)$ were i.i.d. random variables of normal distribution with mean 0 and standard deviation σ. In the *Wiener* case for the models M1–M6, the errors ϵ_{ij} were i.i.d. standard Wiener processes with dispersion parameter σ^2. The values of the parameter σ were $\sigma_1 = 0.2/25$, $\sigma_2 = 1/25$, $\sigma_3 = 1.8/25$, $\sigma_4 = 2.6/25$, $\sigma_5 = 3.4/25$, $\sigma_6 = 4.2/25$ and $\sigma_7 = 5/25$.

Figures 1 and 2 in the Supplementary Materials depict simulated examples corresponding with models M1–M6 with $n_i = 30$, $i = 1, 2, 3$, $\sigma_1 = 0.2/25$ in the normal and the Wiener cases.

For each pair (Mi, σ_j), $i, j = 1, \ldots, 7$ in both cases, the fifteen tests were applied to the three generated functional samples. Their p-values were then recorded. When the p-values are smaller than the nominal significance level $\alpha = 0.05$, the null hypothesis (1) is rejected. The aforementioned process was repeated $N = 100$ times. For each case, the empirical sizes or powers of the tests were then computed as the proportional

Table 2 Empirical sizes (as percentages) of all tests for the one-way ANOVA problem obtained in simulation S1 with $n_i = 10, i = 1, 2, 3$

Data set	Test														
	FP	LH	R	P	W	CH	CS	L^2N	L^2B	L^2b	FN	FB	Fb	GPF	FL
CBF	8	8	20	6	6	4	4	7	9	6	5	8	2	6	26
ChlorineConcentration	8	3	27	2	2	4	5	7	7	6	5	6	3	7	33
CinC_ECG_torso	4	7	24	4	5	4	5	6	6	6	5	6	2	4	39
Cricket_X	7	5	27	4	4	5	7	10	11	8	8	10	5	10	36
Cricket_Y	7	5	24	4	4	3	4	4	5	4	4	4	3	4	35
DiatomSizeReduction	4	3	19	2	4	2	4	4	4	4	3	4	1	3	31
Face (all)	4	3	19	3	1	3	3	4	5	4	3	5	2	4	29
Fish	8	5	31	4	8	4	4	6	9	6	5	5	2	8	29
Haptics	8	5	16	2	5	4	4	5	7	4	5	5	1	4	40
Non-Invasive Thorax1	4	9	31	7	9	4	6	7	10	6	6	8	2	6	39
Non-Invasive Thorax2	3	3	27	3	3	5	7	8	10	9	7	8	4	8	42
OSU Leaf	8	2	21	5	3	4	4	4	5	4	4	4	2	5	25
Plane	7	7	25	0	2	2	5	6	8	5	4	5	2	7	31
StarLightCurves	7	4	22	3	2	5	6	7	7	6	6	7	5	4	39
Swedish Leaf	6	9	28	6	7	3	3	7	9	3	5	6	0	5	28
Symbols	5	7	28	3	5	1	2	3	4	1	1	3	1	5	46
Synthetic Control	8	7	11	6	7	3	2	5	9	3	4	9	0	5	19
Trace	1	5	31	2	4	4	6	6	11	7	6	6	2	8	40
uWaveGestureLibrary_X	3	1	20	4	1	6	6	6	6	6	6	6	3	7	35
uWaveGestureLibrary_Y	7	0	25	2	2	3	4	6	8	5	5	6	1	6	31
uWaveGestureLibrary_Z	7	2	23	4	2	4	4	5	5	4	4	5	1	5	34
Mean	6	5	24	4	4	4	5	6	7	5	5	6	2	6	34

number of rejections (out of $N = 100$ replications) based on the calculated p-value (see Tables 6–17 in the Supplementary Materials).

3.2 Results

It is not easy to draw conclusions about the behavior of all of the tests for (1) being compared by means of the simulations given in Sect. 3.1. For this reason, we present a detailed statistical comparison to identify differences between the tests in Sect. 3.3. In this section, we restrict ourselves to providing simple and immediate insights obtained from Tables 2, 3 and Tables 2–17 in the Supplementary Materials.

In simulation S1 and in models M1 and M2 (where H_0 is true) all tests except the R and FL tests provide satisfactory results in almost all of the cases considered. From Table 2 and Tables 2, 3, 6–8, 12–14 in the Supplementary Materials, we immediately observe that the R and FL tests are the most liberal of all the tests, with empirical sizes ranging between 10 and 46 %, 4 and 46 % respectively. Hence it seems that the

Table 3 Empirical powers (as percentages) of all tests for the one-way ANOVA problem obtained in simulation S2 with $n_i = 10, i = 1, 2, 3$

Data set	Test														
	FP	LH	R	P	W	CH	CS	L^2N	L^2B	L^2b	FN	FB	Fb	GPF	FL
Adiac	85	80	97	49	81	86	86	87	89	87	87	87	78	93	100
CBF	100	100	100	98	100	97	97	98	99	98	97	98	86	100	99
ChlorineConcentration	4	10	25	11	9	5	5	6	8	5	6	6	2	7	47
CinC_ECG_torso	16	17	51	6	12	8	9	11	15	9	9	12	6	10	45
Cricket_X	36	22	60	14	24	25	27	32	38	30	29	33	20	39	74
Cricket_Y	66	31	76	23	34	62	60	66	70	62	64	66	51	78	88
DiatomSizeReduction	85	97	100	35	97	58	57	63	64	60	58	59	52	91	97
Face (all)	80	80	89	35	75	82	79	84	89	82	84	89	70	95	100
Face (four)	100	93	100	26	86	96	96	97	100	96	97	98	94	100	100
Fish	77	69	94	23	63	71	72	72	75	72	72	72	65	85	100
Haptics	37	12	34	9	12	23	24	31	33	24	27	31	14	37	67
Mallat	100	94	100	41	88	100	100	100	100	100	100	100	100	100	100
MedicalImages	90	46	80	24	44	82	82	84	86	83	84	85	72	84	94
Non-Invasive Thorax1	100	93	99	32	88	99	98	99	99	98	99	99	97	99	100
Non-Invasive Thorax2	99	97	100	38	95	100	100	100	100	100	100	100	100	100	100
OSU Leaf	40	9	47	13	13	25	24	31	35	24	29	32	17	34	75
Plane	100	100	100	24	100	100	100	100	100	100	100	100	99	100	100
StarLightCurves	86	77	92	40	79	85	87	87	90	87	86	87	83	90	94
Swedish Leaf	96	79	98	35	78	93	90	94	94	93	93	94	84	99	99
Symbols	98	91	96	37	90	97	97	97	97	97	97	97	95	99	100
Synthetic Control	98	91	96	89	91	93	89	94	98	91	93	98	85	100	100
Trace	70	84	95	32	82	69	69	69	69	69	69	69	69	85	90
uWaveGestureLibrary_X	96	37	69	25	39	94	94	94	95	94	94	95	91	95	97
uWaveGestureLibrary_Y	88	33	76	25	34	91	92	93	93	92	92	92	90	93	98
uWaveGestureLibrary_Z	91	41	74	21	43	89	90	91	95	90	90	91	85	94	92
Mean	78	63	82	32	62	73	73	75	77	74	74	76	68	80	90

significance levels of the R and FL tests are not less than or equal to 5%. However, it also seems that with increasing sample sizes, the empirical sizes of the FL test generally decrease, but we can not say the same for the R test. The other tests perform much better than the R and FL tests. In simulation S1, the L^2B, FB, L^2N, FP and GPF tests are slightly more liberal than all of the remaining tests except the R and FL tests. The Fb test is the most conservative. In models M1 and M2 in the normal case, the LH, W, P, L^2B, FP and FB tests are slightly more liberal than the GPF and L^2N tests, and the CH, CS, L^2b, FN and Fb tests are the most conservative. In models M1 and M2 in the Wiener case, the L^2B, GPF, L^2N, CS and L^2b tests are slightly more liberal than the other tests, and the P, CH and Fb tests are usually the most conservative (the LH and W tests are also conservative when $n_i = 30, i = 1, 2, 3$).

Information about the empirical powers of the tests is obtained from simulation S2 and models M3–M7 (see Table 3; Tables 4, 5, 9–11, 15–17 in the Supplementary Materials). The R and FL tests seem to be ones of the most powerful tests, but this may be connected with the unacceptable high empirical sizes of those tests. In simulation S2, the GPF, FP, L^2B and FB tests have slightly higher powers than the L^2N, L^2b and FN tests, and these all have higher powers than the other tests. The P test has the smallest empirical power. In models M3–M7 in the normal case, the FP, L^2B and FB tests have the highest powers. The remaining tests have smaller powers than those tests, and the Fb and P tests have the smallest empirical powers. In models M3–M6 in the Wiener case, the situation is changing. Here, LH, W, P (for $n_i = 20, 30$, $i = 1, 2, 3$) and GPF tests are the most powerful tests. Nevertheless, the Fb test has still the smallest empirical power.

With increasing sample sizes, the empirical sizes of the tests generally become better or are at the same level in terms of size control, and the empirical powers of the tests generally increase.

The powers of the tests for (1) depend strongly on the error parameters. With an increase in the dispersion parameter, the empirical powers of the tests generally become smaller. All of the tests (except eventually the P test in the normal case or when $n_i = 10$, $i = 1, 2, 3$) perform very well for σ_1 and well for σ_2 and σ_3. However, for σ_i, $i = 5, 6, 7$, the CH, CS, L^2N, L^2b, FN and Fb tests in each case and the FP, L^2B and FB tests in the Wiener case often perform badly or even very badly. All tests depend on the dispersion parameter, but some of them to a much greater degree than others.

In model M7 the functions are in fact constant, so this model is an example in which the functional approach is unnecessarily complicated. We include this case to compare the functional tests with the classical ANOVA F-test for real variables, which would be more appropriate here. In model M7, the classical ANOVA F-test gives 97 % rejections of the null hypothesis (1) at a significance level $\alpha = 0.05$ for σ_7 and $n_i = 10, i = 1, 2, 3$, and 100 % rejections in all other cases. Although the functional tests perform generally worse than the classical ANOVA F-test for real variables, most of them (especially the FP, L^2B and FB tests) perform quite well for small and moderate values of the dispersion parameter.

We can also observe that in simulation S2 and models M3–M7 in the normal case the CH test performs slightly better than the CS test. The reason for this is perhaps that the homoscedastic assumption appears reasonable for those simulation data. The opposite situation holds in models M3–M6 in the Wiener case.

3.3 Statistical comparison of tests

To identify differences between the tests, we present a detailed statistical comparison. We test the null hypothesis that all tests perform the same and the observed differences are merely random. We used the Iman and Davenport (1980) test, which is a nonparametric equivalent of ANOVA. The F-test is recommended because it is less conservative than other tests (Looney 1998; Demšar 2006). We perform this test separately for the results of the simulation S1, the simulation S2, the models M1 and M2 in each case, the models M3–M7 in the normal case, the models M3–M6 in the Wiener

case, and separately for $n_i = 10, 20, 30, i = 1, 2, 3$ in each of these simulation scenarios. The F-test ranks the tests for each data set or pair (Ms, σ_t), $s, t = 1, \ldots, 7$, separately, the tests with the smallest rates of rejection of the null hypothesis receiving a rank of 1, the tests with the second smallest rates of rejection a rank of 2, and so on (in case of ties average ranks are assigned). Let R_{ij} be the rank of the jth of K tests on the ith of N data sets or pairs (Ms, σ_t), and $R_j = \frac{1}{N}\sum_{i=1}^{N} R_{ij}$. The test compares the mean ranks of methods and is based on the statistic

$$S' = \frac{(N-1)S}{N(K-1) - S},$$

where

$$S = \frac{12N}{K(K+1)} \sum_{i=1}^{K} R_i^2 - 3N(K+1)$$

is the Friedman statistic, which has the F distribution with $K-1$ and $(K-1)(N-1)$ degrees of freedom. The p-values from this test are less than 2.2E−16 for all simulation scenarios described in Sect. 3.1 except simulations S1 with $n_i = 30, i = 1, 2, 3$ (p-value = 3.005E−10) and S2 with $n_i = 20, 30, i = 1, 2, 3$ (p-value = 3.281E−12, 9.721E−06 respectively). We can therefore proceed with the post hoc tests to detect significant pairwise differences among all of the tests. A set of pairwise comparisons can be associated with a set of hypotheses. Any of the post hoc tests which can be applied to non-parametric tests work over a family of hypotheses. The test statistic for comparing the ith and jth tests is

$$Z = \frac{R_i - R_j}{\sqrt{\frac{K(K+1)}{6N}}}.$$

This statistic is asymptotically normal with zero mean and unit variance. When comparing multiple algorithms, to retain an overall significance level α, one has to adjust the value of α for each post hoc comparison. There are various methods for this. A simple method is to use Bonferroni correction. There are $m = K(K-1)/2$ comparisons, therefore Bonferroni correction sets the significance level of each comparison to α/m. Demšar (2006) recommends the procedure of Nemenyi (1963), which is based on this correction. The Nemenyi test is similar to the Tukey test for ANOVA and is used when all tests are compared with each other. The performance of two tests is significantly different at the experimentwise error rate α if

$$|R_i - R_j| > q(\alpha, K, \infty) \left(\frac{K(K+1)}{12N}\right)^{1/2}, \tag{7}$$

$i = 1, \ldots, K-1$, $j = i+1, \ldots, K$, where the values of $q(\alpha, K, \infty)$ are based on the Studentized range statistic (Hollander and Wolfe 1973; Demšar 2006). The results of multiple comparisons are given in Table 4 and Tables 18 and 19 in the

Supplementary Materials. Those tests that are connected by a sequence of letters have average ranks that are not significantly different from one another. For example, when $n_i = 10, i = 1, 2, 3$ for the empirical size, we obtained six homogeneous groups of tests for simulation S1, and five homogeneous groups of tests for models M1 and M2 in each case, and for the empirical power, seven homogeneous groups of tests for simulation S2 and models M3–M6 in the Wiener case, and six homogeneous groups of tests for models M3–M7 in the normal case (see Table 4). For the empirical size, the best tests are in the last group, while for the empirical power, the best tests are in the first group.

4 Illustrative examples

In this section, we apply the tests to two real-data examples, using Canadian temperature data and orthosis data. These data sets are commonly used to illustrate the use of statistical methods for real functional data (see Abramovich et al. 2004; Ramsay and Silverman 2005; Zhang 2013; Zhang and Liang 2014). The Canadian temperature data are available in the R package fda, and the orthosis data can be downloaded from "http://www.stat.nus.edu.sg/~zhangjt/books/Chapman/FANOVA.htm".

4.1 Canadian temperature data

The Canadian temperature data are the daily temperature records of 35 Canadian weather stations over a year (365 days). Fifteen of the weather stations are in Western Canada, another fifteen in Eastern Canada, and the remaining five in Northern Canada. Panels (a)–(c) of Fig. 1 present the raw Canadian temperature curves for these 35 weather stations. From Fig. 1, it can be seen that the temperatures at the Eastern and Western weather stations are generally higher than those at the Northern weather stations. The reason for this is probably that the Northern stations are located at higher latitudes. We would like to check statistically whether location has an effect on the mean temperature curves of the Eastern, Western and Northern weather stations. This problem was considered by Zhang (2013), and it is equivalent to the one-way ANOVA problem for functional data with $k = 3$. To solve this problem, we applied all of the tests under consideration, obtaining the results given in Table 5 (the estimates of the parameters of the L^2N, L^2B, FN, FB and GPF tests are given in Table 20 in the Supplementary Materials). The p-values of the first fourteen tests are less than the significance level 0.05, and the value of the test statistic of the FL test is greater than the critical value, hence it can be concluded that location has an effect on the mean temperature curves of the Eastern, Western and Northern weather stations.

4.2 Orthosis data

As reported by Abramovich et al. (2004), the orthosis data were acquired and computed in an experiment by Dr. Amarantini David and Dr. Martin Luc (Laboratoire Sport et

Table 4 Results of the Nemenyi post hoc test for the results of the simulations S1 and S2, the models M1 and M2, and the models M3–M7 with $n_i = 10, i = 1, 2, 3$

Test	Ranks mean	Homogeneous groups	Test	Ranks mean	Homogeneous groups
Empirical size (S1)			*Empirical power (S2)*		
FL	14.93	a	FL	13.74	a
R	14.07	a	GPF	12.02	a b
L^2B	11.79	a b	R	11.02	a b c
FB	9.31	b c	L^2B	11.00	a b c
L^2N	9.17	b c d	FP	10.18	a b c d
FP	9.07	b c d	FB	9.48	a b c d
GPF	8.48	b c d e	L^2N	9.04	b c d e
L^2b	7.36	b c d e	FN	7.82	b c d e f
LH	6.86	c d e	L^2b	7.34	c d e f
FN	6.38	c d e f	CH	6.52	d e f
CS	5.98	c d e f	CS	6.40	d e f
W	5.62	c d e f	LH	5.08	e f g
P	4.62	d e f	W	4.74	f g
CH	4.21	e f	Fb	3.78	f g
Fb	2.17	f	P	1.84	g
Empirical size (M1, M2 in normal case)			*Empirical power (M3–M7 in normal case)*		
R	14.82	a	FL	12.64	a
FL	14.18	a	FP	11.51	a b
LH	12.43	a	L^2B	11.37	a b
W	11.32	a b	FB	10.80	a b
P	10.32	a b	GPF	9.61	a b c
L^2B	10.07	a b	R	9.03	a b c d
FP	9.50	a b c	L^2N	8.73	b c d
FB	9.11	a b c d	FN	7.97	b c d e
GPF	6.18	b c d e	CH	7.10	c d e
L^2N	4.29	c d e	L^2b	6.81	c d e
CH	3.86	c d e	LH	6.00	c d e f
FN	3.86	c d e	CS	5.87	d e f
L^2b	3.50	d e	W	5.09	e f
CS	3.29	e	Fb	4.57	e f
Fb	3.29	e	P	2.89	f

Performance Motrice, EA 597, UFRAPS, Grenoble University, France). They investigated how muscle redundancy could be appropriately used to cope with an external perturbation while complying with the mechanical requirements related either to balance control and/or minimum energy expenditure. In the experiment, seven young male volunteers wore a spring-loaded orthosis of adjustable stiffness under the following

Table 4 continued

Test	Ranks mean	Homogeneous groups	Test	Ranks mean	Homogeneous groups
Empirical size (M1, M2 in Wiener case)			*Empirical power (M3–M6 in Wiener case)*		
R	14.75	a	FL	12.91	a
FL	14.25	a	R	12.57	a
L^2B	12.25	a b	GPF	11.18	a b
GPF	10.14	a b c	LH	10.34	a b c
L^2b	9.89	a b c	W	10.00	a b c d
CS	9.54	a b c	L^2B	9.23	a b c d e
L^2N	9.43	a b c	L^2N	8.05	b c d e f
LH	8.50	b c d	L^2b	7.96	b c d e f
FB	6.36	c d e	CS	6.93	c d e f g
FN	6.11	c d e	P	6.14	d e f g
W	5.82	c d e	FP	6.05	d e f g
FP	5.14	c d e	FB	5.82	e f g
P	3.21	d e	FN	5.11	f g
CH	3.07	d e	CH	4.64	f g
Fb	1.54	e	Fb	3.05	g

The critical values of this test given in (7) for the results of the simulation S1, the simulation S2, the models M1 and M2 in each case, the models M3–M7 in the normal case, and the models M3–M6 in the Wiener case with $n_i = 10$, $i = 1, 2, 3$ are equal to 4.680342, 4.289605, 5.732225, 3.625378 and 4.053295 respectively

four experimental conditions: a control condition (without orthosis); an orthosis condition (with orthosis); and two spring conditions (with spring 1 or with spring 2) in which stepping-in-place was perturbed by fitting a spring-loaded orthosis onto the right knee joint. All volunteers tried all four conditions 10 times for 20 s each. To avoid possible perturbations in the initial and final parts of the experiment, only the central 10 s were used in the study. The resultant moment of force at the knee was derived by means of body segment kinematics recorded with a sampling frequency of 200 Hz. For each stepping-in-place replication, the resultant moment was computed at 256 time points, equally spaced and scaled to the interval [0, 1] so that a time interval corresponded to an individual gait cycle.

Similarly to Zhang and Liang (2014), we use only the first volunteer's orthosis data under the four experimental conditions for illustrative purposes. The raw orthosis curves of the first volunteer under the four experimental conditions are given in Panels (a)–(d) of Fig. 2. We are interested in testing whether the mean orthosis curves of the first volunteer are different under the four experimental conditions. This is a one-way ANOVA problem for functional data with $k = 4$. Zhang and Liang (2014) used the pointwise F-test for this problem. This test suggests that the mean orthosis curves of the first volunteer under the four experimental conditions are not the same, but they may be the same over the interval [0.8, 1], i.e. at the last stage of the experiment. The results of all of the tests under consideration are presented in Table 5 (the estimates of the parameters of the L^2N, L^2B, FN, FB and GPF tests are given in Table 20 in

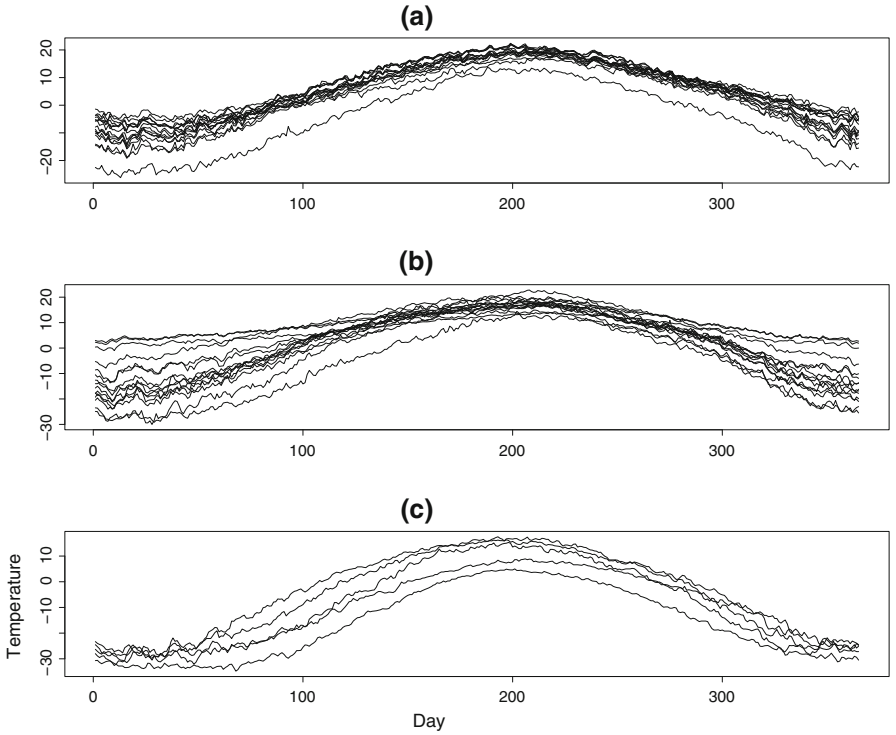

Fig. 1 Canadian temperature data for **a** fifteen Eastern weather stations, **b** fifteen Western weather stations, and **c** five Northern weather stations

the Supplementary Materials). All of the tests reject the null hypothesis that the mean orthosis curves of the first volunteer under the four experimental conditions over [0, 1] do not differ significantly. However, most of the tests accept the equality of these curves over [0.8, 1]. The exceptions are the MANOVA-based tests and the FL test. Hence we see that the decisions suggested by the tests are not always the same in practice.

5 Conclusions

In this paper we have presented comprehensive simulation studies to compare existing tests and new tests for the one-way ANOVA problem (1) for functional data. A large part of the simulations was based on real labeled time series data sets, and for this reason they describe the size and the power of the tests better than simulations based on artificial data.

The simulations suggest that the tests do not perform equally well. Moreover, there is no single test that performs best. When the functions are observed on a moderate or large grid of design time points, the GPF, L^2B, FP and FB tests seem to perform best, although the FP test can be time-consuming. The situation changes when the functions are observed on a short grid of design time points. Here, depending on the

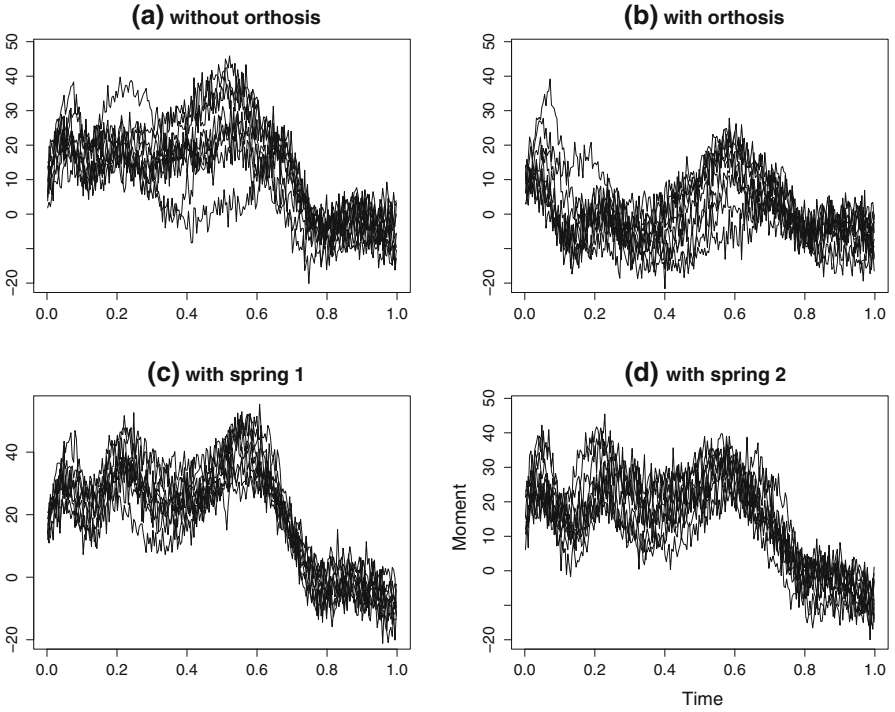

Fig. 2 **a**–**d** Raw orthosis curves for the first volunteer under the four experimental conditions

structure of the functional data, the FP test or the LH, W and P tests seem to perform best. Moreover, the FP test is also very fast in such a case.

Finally, it is worth to mention that the results of performance of the R and FL tests confirm that the investigation of the empirical size of a test is very important and should not be omitted.

6 Supplementary Materials

Supplementary Materials contain the R codes of the programs, which allow to perform the tests for the one-way ANOVA problem for functional data considered in the paper, and the tables, which contain empirical sizes and powers of those tests obtained in remaining simulations S1 and S2 respectively, and in models M1–M2 and M3–M7 respectively; the results of multiple comparisons of tests in simulations S1 and S2 and models M1–M7 with $n_i = 20, 30, i = 1, 2, 3$ (see Sect. 3.3); the estimates of the parameters of the L^2N, L^2B, FN, FB and GPF tests for the Canadian temperature data and the orthosis data for $[a, b] = [0, 1]$ and $[a, b] = [0.8, 1]$. We also present the figures, which depict simulated examples corresponding with models M1–M6.

Table 5 Values of test statistics and p-values of all tests for the Canadian temperature data and the orthosis data for $[a, b] = [0, 1]$ and $[a, b] = [0.8, 1]$

Test	Canadian temp. data		Orthosis data [0, 1]		Orthosis data [0.8, 1]	
	Test stat.	p-value	Test stat.	p-value	Test stat.	p-value
FP	16.312	0	24.063	0	1.1142	0.357
LH	8.3499	3.1E−13	20.942	2.0E−15	0.7786	0.005
R	5.7812	1.1E−09	14.863	8.3E−11	0.7004	0.000
P	1.5723	2.3E−13	2.2715	8.0E−10	0.4844	0.020
W	0.0413	2.5E−13	0.0052	6.6E−13	0.5455	0.011
CH	2.06E6	0	3.02E6	0	1.77E4	0.328
CS	2.06E6	0	3.02E6	0	1.77E4	0.319
L^2N	3.04E5	1.5E−10	7.55E5	0	4425.4	0.352
L^2B	3.04E5	2.7E−11	7.55E5	0	4425.4	0.337
L^2b	3.04E5	0	7.55E5	0	4425.4	0.342
FN	16.210	2.0E−07	19.933	0	1.1098	0.363
FB	16.210	1.3E−07	19.933	0	1.1098	0.362
Fb	16.210	0	19.933	0	1.1098	0.422
GPF	17.231	4.8E−13	17.061	0	1.1343	0.380
FL	189.66	–	124.53	–	6.3156	–

The critical values of the FL test are equal to 3.843317 ($T^* = 100$), 3.843317 ($T^* = 100$) and 3.820904 ($T^* = 51$) respectively (see Table 1 in the Supplementary Materials)

Appendix

Proof of Proposition 1 Let $\{\phi_l\}$ be an orthonormal basis of the space $L_2(T)$ and $X_{ij}(t) = \mathbf{c}'_{ij}\boldsymbol{\phi}(t)$ for all $t \in T$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, where $\mathbf{c}_{ij} = (c_{ij0}, c_{ij1}, \ldots, c_{ijK})'$ and $\boldsymbol{\phi}(t) = (\phi_0(t), \phi_1(t), \ldots, \phi_K(t))'$. By (5), we have

$$\bar{X}(t) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{c}'_{ij}\boldsymbol{\phi}(t)$$

and

$$\bar{X}_i(t) = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{c}'_{ij}\boldsymbol{\phi}(t).$$

The definition of the norm of the space $L_2(T)$ implies

$$||\bar{X}_i - \bar{X}||_2^2 = \int_T (\bar{X}_i(t) - \bar{X}(t))^2 dt.$$

From the basis function representation of $X_{ij}(t)$, it follows that

$$||\bar{X}_i - \bar{X}||_2^2 = \int_T \left(\frac{1}{n_i} \sum_{m=1}^{n_i} \mathbf{c}'_{im} \boldsymbol{\phi}(t) - \frac{1}{n} \sum_{j=1}^{k} \sum_{p=1}^{n_j} \mathbf{c}'_{jp} \boldsymbol{\phi}(t) \right)^2 dt.$$

Hence $||\bar{X}_i - \bar{X}||_2^2$ is equal to

$$\left(\frac{1}{n_i} \sum_{m=1}^{n_i} \mathbf{c}'_{im} - \frac{1}{n} \sum_{j=1}^{k} \sum_{p=1}^{n_j} \mathbf{c}'_{jp} \right) \int_T \boldsymbol{\phi}(t) \boldsymbol{\phi}'(t) dt \left(\frac{1}{n_i} \sum_{s=1}^{n_i} \mathbf{c}_{is} - \frac{1}{n} \sum_{t=1}^{k} \sum_{v=1}^{n_t} \mathbf{c}_{tv} \right).$$

Since $\{\phi_l\}$ is an orthonormal basis, $\int_T \boldsymbol{\phi}(t) \boldsymbol{\phi}'(t) dt$ is equal to the identity matrix of size $K + 1$. Therefore

$$\begin{aligned} ||\bar{X}_i - \bar{X}||_2^2 = {} & \frac{1}{n_i^2} \sum_{m=1}^{n_i} \sum_{s=1}^{n_i} \mathbf{c}'_{im} \mathbf{c}_{is} - \frac{2}{n n_i} \sum_{m=1}^{n_i} \sum_{t=1}^{k} \sum_{v=1}^{n_t} \mathbf{c}'_{im} \mathbf{c}_{tv} \\ & + \frac{1}{n^2} \sum_{j=1}^{k} \sum_{p=1}^{n_j} \sum_{t=1}^{k} \sum_{v=1}^{n_t} \mathbf{c}'_{jp} \mathbf{c}_{tv}. \end{aligned}$$

Thus since $n_1 + \cdots + n_k = n$, $\sum_{i=1}^{k} n_i ||\bar{X}_i - \bar{X}||_2^2 = a - b$, where a and b are given in Proposition 1. The denominator of F given by (6) may be handled in much the same way, which completes the proof. □

Proof of Lemma 1 Let $\mathbf{C}_i = (\mathbf{c}_{i1}, \ldots, \mathbf{c}_{in_i})$, $i = 1, \ldots, k$. For $i, j = 1, \ldots, k$, we have

$$\mathbf{C}'_i \mathbf{C}_j = \begin{pmatrix} \mathbf{c}'_{i1}\mathbf{c}_{j1} & \mathbf{c}'_{i1}\mathbf{c}_{j2} & \ldots & \mathbf{c}'_{i1}\mathbf{c}_{jn_j} \\ \mathbf{c}'_{i2}\mathbf{c}_{j1} & \mathbf{c}'_{i2}\mathbf{c}_{j2} & \ldots & \mathbf{c}'_{i2}\mathbf{c}_{jn_j} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{c}'_{in_i}\mathbf{c}_{j1} & \mathbf{c}'_{in_i}\mathbf{c}_{j2} & \ldots & \mathbf{c}'_{in_i}\mathbf{c}_{jn_j} \end{pmatrix}.$$

Hence the sums $\sum_{m=1}^{n_i} \sum_{s=1}^{n_i} \mathbf{c}'_{im}\mathbf{c}_{is}$, $\sum_{m=1}^{n_i} \sum_{v=1}^{n_t} \mathbf{c}'_{im}\mathbf{c}_{tv}$, $\sum_{j=1}^{n_i} \mathbf{c}'_{ij}\mathbf{c}_{ij}$ are equal to the sum of entries of the matrix $\mathbf{C}'_i\mathbf{C}_i$, the sum of entries of the matrix $\mathbf{C}'_i\mathbf{C}_t$ and the trace of the matrix $\mathbf{C}'_i\mathbf{C}_i$ respectively, i.e. they are equal to $\mathbf{1}'_{n_i}\mathbf{C}'_i\mathbf{C}_i\mathbf{1}_{n_i}$, $\mathbf{1}'_{n_i}\mathbf{C}'_i\mathbf{C}_t\mathbf{1}_{n_t}$ and $\mathrm{tr}(\mathbf{C}'_i\mathbf{C}_i)$, respectively. This finishes the proof. □

References

Abramovich F, Antoniadis A, Sapatinas T, Vidakovic B (2004) Optimal testing in a fixed-effects functional analysis of variance model. Int J Wavelets Multiresolut Inf Process 2:323–349

Anderson TW (2003) An introduction to multivariate statistical analysis, 3rd edn. Wiley, London

Benhenni K, Ferraty F, Rachdi M, Vieu P (2007) Local smoothing regression with functional data. Comput Stat 22:353–369

Berrendero JR, Justel A, Svarc M (2011) Principal components for multivariate functional data. Comput Stat Data Anal 55:2619–2634

Bobelyn E, Serban AS, Nicu M, Lammertyn J, Nicolai BM, Saeys W (2010) Postharvest quality of apple predicted by NIR-spectroscopy: study of the effect of biological variability on spectra and model performance. Postharvest Biol Technol 55:133–143

Boente G, Fraiman R (2000) Kernel-based functional principal components. Ann Stat 20:655–674

Cai T, Hall P (2006) Prediction in functional linear regression. Ann Stat 34:2159–2179

Chiou JM, Müller HG (2007) Diagnostics for functional regression via residual processes. Comput Stat Data Anal 15:4849–4863

Cuevas A (2014) A partial overview of the theory of statistics with functional data. J Stat Plan Inference 147:1–23

Cuevas A, Febrero M, Fraiman R (2002) Linear functional regression: the case of fixed design and functional response. Can J Stat 30:285–300

Cuevas A, Febrero M, Fraiman R (2004) An anova test for functional data. Comput Stat Data Anal 47:111–122

Davidian M, Lin X, Wang J (2004) Introduction: emerging issues in longitudinal and functional data analysis. Stat Sin 14:613–614

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Fan J, Lin SK (1998) Test of significance when data are curves. J Am Stat Assoc 93:1007–1021

Faraway J (1997) Regression analysis for a functional response. Technometrics 39:254–261

Ferraty F, Vieu P (2006) Nonparametric functional data analysis: theory and practice. Springer, New York

Gao HO (2007) Day of week effects on diurnal ozone/NOx cycles and transportation emissions in Southern California. Transp Res Part D 12:292–305

Górecki T, Krzyśko M, Waszak Ł (2014) Functional discriminant coordinates. Commun Stat Theory Methods 43:1013–1025

Hollander M, Wolfe DA (1973) Nonparametric statistical methods. Wiley, New York

Horváth L, Kokoszka P (2012) Inference for functional data with applications. Springer, New York

Iman RL, Davenport JM (1980) Approximations of the critical region of the Friedman statistic. Commun. Stat. Theory Methods 9:571–595

James GM, Hastie TJ (2001) Functional linear discriminant analysis for irregularly sampled curves. J R Stat Soc Ser B (Stat Methodol) 63:533–550

Keogh E, Zhu Q, Hu B, Hao Y, Xi X, Wei L, Ratanamahatana CA (2011) The UCR Time Series Classification/Clustering Homepage. http://www.cs.ucr.edu/~eamonn/time_series_data/

Krzyśko M, Waszak Ł (2013) Canonical correlation analysis for functional data. Biometr Lett 50:95–105

Laukaitis A, Račkauskas A (2005) Functional data analysis for clients segmentation tasks. Eur J Oper Res 163:210–216

Leurgans SE, Moyeed RA, Silverman BW (1993) Canonical correlation analysis when the data are curves. J R Stat Soc Ser B (Stat Methodol) 55:725–740

Long W, Li N, Wang H, Cheng S (2012) Impact of US financial crisis on different countries: based on the method of functional analysis of variance. Procedia Comput Sci 9:1292–1298

Looney SW (1998) A statistical technique for comparing the accuracies of several classifiers. Pattern Recogn Lett 8:5–9

Martínez-Camblor P, Corral N (2011) Repeated measures analysis for functional data. Comput Stat Data Anal 55:3244–3256

Nemenyi PB (1963) Distribution-free multiple comparisons. Dissertation, Princeton University

Preda C, Saporta G, Lévéder C (2007) PLS classification of functional data. Comput Stat 22:223–235

Ramsay JO, Hooker G, Graves S (2009) Functional data analysis with R and MATLAB. Springer, Berlin

Ramsay JO, Silverman BW (2002) Applied functional data analysis: methods and case studies. Springer, New York

Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, New York

Shen Q, Faraway J (2004) An F test for linear models with functional responses. Stat Sin 14:1239–1257
Tarrío-Saavedra J, Naya S, Francisco-Fernández M, Artiaga R, Lopez-Beceiro J (2011) Application of functional ANOVA to the study of thermal stability of micronano silica epoxy composites. Chemometr Intell Lab Syst 105:114–124
Tokushige S, Yadohisa H, Inada K (2007) Crisp and fuzzy k-means clustering algorithms for multivariate functional data. Comput Stat 22:1–16
Valderrama MJ (2007) An overview to modelling functional data. Comput Stat 22:331–334
Yamamoto M, Terada Y (2014) Functional factorial K-means analysis. Comput Stat Data Anal 79:133–148
Zhang JT (2011) Statistical inferences for linear models with functional responses. Stat Sin 21:1431–1451
Zhang JT (2013) Analysis of variance for functional data. Chapman & Hall, London
Zhang JT, Chen JW (2007) Statistical inferences for functional data. Ann Stat 35:1052–1079
Zhang JT, Liang X (2014) One-way ANOVA for functional data via globalizing the pointwise F-test. Scand J Stat 41:51–71
Zhao X, Marron JS, Wells MT (2004) The functional data analysis view of longitudinal data. Stat Sin 14:789–808